An algebra of variant loci

Johannes Köster[1,2,3], Sven Rahmann[4]

[1]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute and Harvard School of Public Health
[2]Department of Medical Oncology, Dana-Farber Cancer Institute and Harvard Medical School
[3]Center for Functional Cancer Epigenetics, Dana-Farber Cancer Institute
[4]Genome Informatics, Institute of Human Genetics, University of Duisburg-Essen

Introduction

Variant calling on NGS data often entails filtering samples against each other to e.g.

- detect de-novo mutations (child vs. parents, tumor vs. normal),
- eliminate sequencing artifacts.

This can be formulated as set operations, e.g.

$$V_A \setminus (V_B \cup V_C)$$

with V_A, V_B and V_C being the true variant loci of sample A, B and C. Of course, these sets are unknown. The state of the art is to call variants of each sample, and perform set-based filtering afterwards.

This gives rise to three problems:

Insufficient evidence problem The filtering fails if the coverage is too low.

N + 1 problem Calling samples in groups helps with the insufficient evidence problem. But later addition of new samples leads to redundant calculations.

FDR problem The obtained variant qualities do not reflect the filtering. This makes controlling the false discovery rate (FDR) difficult.

Apart from specialized solutions for Tumor/Normal pairs, no solution for generic filtering scenarios exists that solves all problems.

We present **Algebraic Variant Calling**, an approach to incorporate the filtering into the calling model. Algebraic Variant Calling solves the insufficient evidence problem and provides intuitive FDR control. In the ALgebraic PArallel CAller (**ALPACA**), we combine Algebraic Variant Calling with a BCF based approach to solve the N+1 problem.

Algebra of variant loci

For a finite set of samples $S = \{s_1, s_2, \dots\}$ with variant loci $V_S = \{V_{s_1}, V_{s_2}, \dots\}$, we define the algebra

$$\mathcal{A}_S = \left(2_S^V \setminus \emptyset, \cup, \setminus, \left(\bigotimes^k\right)_{k\in\mathbb{N}}\right)$$

with the classic set operations union $\cup$ and difference $\setminus$ and a k-relaxed intersection $\bigotimes^k$. The k-relaxed intersection $\bigotimes^k_{s\in S'} V_s$ for subset $S' \subseteq S$ with $|S'| \geq k$ is the set of variant loci common to at least k of the samples in S'.

This allows all kinds of filtering scenarios, e.g.

- Call all variants in any of the samples a, b, c:
$$V_a \cup V_b \cup V_c$$
- Call somatic mutations in e.g. a tumor sample t compared to a healthy normal sample n:
$$V_t \setminus V_n$$
- Call de-novo mutations in a child sample c compared to its parents f, m:
$$V_c \setminus (V_f \cup V_m)$$
- Call somatic mutations in a group of tumors t, t' compared to their normals n, n':
$$(V_t \cup V_{t'}) \setminus (V_n \cup V_{n'})$$
- Do the same in a paired way:
$$(V_t \setminus V_n) \cup (V_{t'} \setminus V_{n'})$$
- Call all variants that are recurrent in at least 3 of the samples a, b, c, d, e:
$$\bigotimes^3_{s\in\{a,b,c,d,e\}} V_s$$

Algebraic variant calling

For any scenario $\phi \in \mathcal{A}_S$, we calculate the posterior probability for the null hypothesis $i \notin \phi$ for each locus i. If $\phi = \bigcup_{s\in S'\subseteq S} V_s$, we calculate

$$\Pr(i \notin \phi \,|\, D_{S,i}) := \Pr(M = 0 \,|\, D_{S',i})$$

e.g. in the usual Bayesian way (dePristo et al. 2011, Li 2010). If $\phi = \phi_1 \setminus \phi_2$, we write

$$\Pr(i \notin \phi \,|\, D_{S,i}) := 1 - \Pr(i \in \phi_1 \,|\, D_{S,i}) \cdot \Pr(i \notin \phi_2 \,|\, D_{S,i}),$$

and $\phi = \phi_1 \cup \phi_2$ leads to

$$\Pr(i \notin \phi \,|\, D_{S,i}) := \Pr(i \notin \phi_1 \,|\, D_{S,i}) \cdot \Pr(i \notin \phi_2 \,|\, D_{S,i}).$$

For the k-relaxed intersecton $\phi = \bigotimes^k_{i=1,2,\dots} \phi_i$ we can calculate $\Pr(i \notin \phi \,|\, D_{S,i})$ with dynamic programming.

Finally, we can approximate ϕ as

$$\phi^*_\alpha := \{i \mid \forall i = 1, 2, \dots, n : \Pr(i \in \phi | D_{S,i}) \leq \alpha\}.$$

Even low coverage evidence for a variant in a filtering sample will affect the resulting posterior.

Controlling FDR

FDR can be controlled to not exceed α^* by setting the threshold

$$\alpha = \max\{\alpha' \in [0, \alpha^*] \mid \overline{FDR}_{\alpha'} \leq \alpha^*\}$$

with

$$\overline{FDR}_\alpha = \frac{1}{|\phi^*_\alpha|} \sum_{i\in\phi^*_\alpha} \Pr(i \notin \phi | D_{S,i}).$$

Since posterior probabilities reflect the given query, controlling FDR becomes easy.

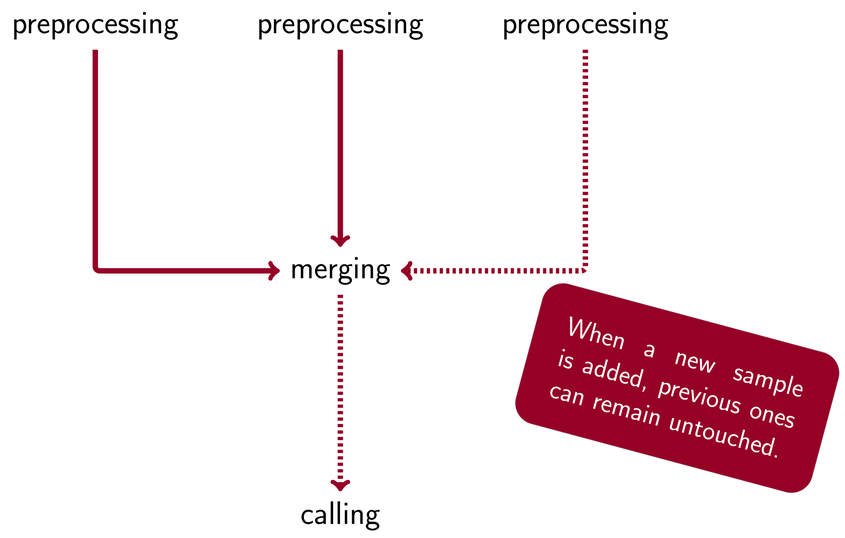
Solving the N+1 problem

The probability $\Pr(M = 0|D_{S,i})$ is calculated from per-sample genotype likelihoods

$$\Pr(D_{s,i}|G = g)$$

with G being the random variable denoting possible genotypes. $D_{s,i}$ is the pileup of read bases of sample s at locus i. The likelihoods are independent of the query formula ϕ, hence:

- Genotype likelihoods for all covered loci can be preprocessed into per-sample BCF files.
- Sample BCF files can be merged into a global BCF, keeping only loci with any non-reference maximum likelihood genotype.
- Calling with different scenarios $\phi \in \mathcal{A}_S$ becomes a matter of seconds.



LaTeX TikZposter